KGOT: UNIFIED KNOWLEDGE GRAPH AND OPTIMAL TRANSPORT PSEUDO-LABELING FOR MOLECULE-PROTEIN INTERACTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting molecule-protein interactions (MPIs) is a fundamental task in computational biology, with crucial applications in drug discovery and molecular function annotation. However, existing MPI models face two major challenges. First, the scarcity of labeled molecule-protein pairs significantly limits model performance, as available datasets capture only a small fraction of biological relevant interactions. Second, most methods rely solely on molecular and protein features, ignoring broader biological context—such as genes, metabolic pathways, and functional annotations—that could provide essential complementary information. To address these limitations, our framework first aggregates diverse biological datasets, including molecular, protein, genes and pathway-level interactions, and then develop an optimal transport-based approach to generate high-quality pseudo-labels for unlabeled molecule-protein pairs, leveraging the underlying distribution of known interactions to guide label assignment. By treating pseudo-labeling as a mechanism for bridging disparate biological modalities, our approach enables the effective use of heterogeneous data to enhance MPI prediction. We evaluate our framework on multiple MPI datasets including virtual screening tasks and protein retrieval tasks, demonstrating substantial improvements over state-of-the-art methods in prediction accuracys and zero shot ability across unseen interactions. Beyond MPI prediction, our approach provides a new paradigm for leveraging diverse biological data sources to tackle problems traditionally constrained by single- or bi-modal learning, paving the way for future advances in computational biology and drug discovery.

1 INTRODUCTION

Molecular and protein representation learning is an increasingly important topic in computational biology and drug discovery (Jumper et al., 2021; Zhou et al., 2023), fueled by the availability of large-scale *unlabeled* datasets (Consortium, 2024; Guo et al., 2024; AlQuraishi, 2019; Nakata et al., 2020). These resources have enabled the development of powerful molecular and protein encoders, which serve as foundational models for various downstream tasks. For example, self-supervised learning approaches (Rives et al., 2019; Zhou et al., 2023) leverage massive sequence and structure databases to capture intricate biochemical properties without requiring explicit supervision. This line of research not only advances standalone molecular/protein modeling but also lays the foundation for pushing the boundaries of computational biology and accelerating medicine and life science discoveries.

Despite these advances, retrieving molecule-protein interactions remains a formidable challenge. A core issue is the scarcity of large-scale labeled datasets for MPIs, due to the experimental complexity and cost of validating interactions. Each new molecule-protein interaction must typically be confirmed via laborious assays, often with regulatory oversight in drug discovery, which severely limits data growth. High-throughput screens are expensive and slow, and even computational docking simulations (Di Nola et al., 1994) are constrained by accuracy and scale. Existing MPI datasets (Chandak et al., 2023; Gao et al., 2023) tend to be small, biased toward specific protein families, or inconsistent in annotations, making it difficult to train deep models that generalize across diverse interactions better than traditional techniques (such as docking and energy scoring (Wang et al., 2020)). If more labeled molecule-protein pairs were available spanning diverse chemistry and tar-

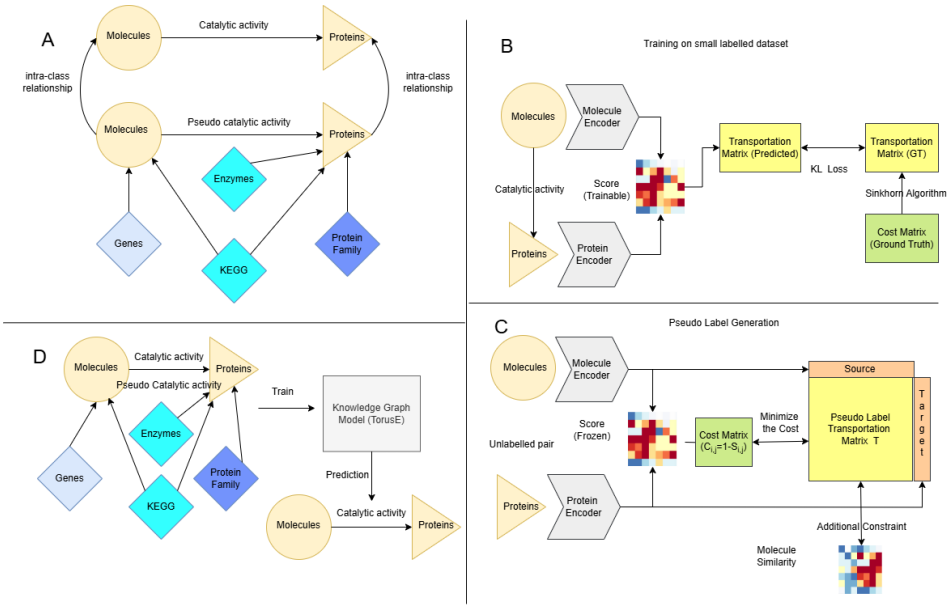

Figure 1: **Overview of KGOT. (A)** *Knowledge integration* **(B)** *Supervised score learning* **(C)** *Pseudo-label generation* **(D)** *KG augmentation & link prediction*

gets, deep learning models could learn richer interaction patterns and capture complex biophysical properties that traditional methods struggle with. Recent studies (Xia et al., 2024) underscore the need for novel frameworks that combine modern representation learning with strategies to cope with limited labels, in order to yield more robust and generalizable interaction predictions.

Another major challenge is the narrow reliance on *bi-modal* data, i.e., using only molecular and protein features while ignoring other relevant biological information. In real-world biology and drug discovery, molecular interactions are influenced by a broad spectrum of factors. For instance, genetic variations can alter protein function and a molecule's binding efficacy; biochemical pathways and networks can modulate the downstream effects of a drug and indicate which proteins are likely involved. These additional modalities provide essential context for understanding interactions beyond what molecular structure or protein sequence alone.

Large-scale biological knowledge graphs (KGs) offer a promising way to integrate these heterogeneous modalities. Resources like PrimeKG (Chandak et al., 2023) aggregate diverse biological entities and relations, but they contain relatively few direct molecule-protein interactions and are not tailored for MPI prediction. Our work aims to bridge this gap by systematically integrating multimodal biological data into the MPI prediction process. By using a biological KG as a structured prior, we contextualize molecule-protein pairs with genetic, biochemical, and phenotypic insights, which can improve prediction robustness and generalization. In our framework, we extract relevant information from diverse sources into a unified graph and then refine the resulting representations through a pseudo-labeling mechanism based on optimal transport, aligning the multimodal knowledge with the MPI prediction task.

To address the data scarcity challenge, we constructed a large-scale *multimodal biological knowledge graph* by integrating six high-quality public datasets. The resulting KG contains over three million relations, encompassing molecules, proteins, genes, pathways, and other biomedical entities. By densely connecting molecule and protein nodes with diverse biological entities, this KG provides a rich relational context that compensates for the lack of direct molecule-protein labels. It also enables the model to capture indirect interaction paths, potentially improving generalization to new interactions.

Building on this knowledge graph, we propose a *pseudo-label generation framework* based on optimal transport to effectively leverage both labeled and unlabeled data. Instead of relying solely on the sparse interaction labels, we formulate pseudo-label assignment as a point-set matching problem:

using OT, we align predicted interaction scores with the underlying biological structure to produce high-confidence pseudo-labels for unlabeled molecule-protein pairs. Concretely, our method first trains an interaction scoring function on top of molecular and protein encoders. We then infer scores for all unlabeled molecule-protein pairs to construct an initial dense score matrix. Next, we apply an OT-based algorithm to assign pseudo-label interaction probabilities by minimizing transport cost between molecule and protein distributions. The resulting pseudo-labeled interactions, combined with ground-truth labels, are used to augment training of the final model. This approach effectively transforms the MPI task into a semi-supervised learning problem, where the model benefits from an expanded training set of confident interaction examples.

In summary, our contributions are threefold:

- We construct a large-scale **multimodal knowledge graph** for MPI prediction by integrating diverse public datasets. This structured graph representation connects molecules and proteins through biological relationships, enabling the model to utilize multimodal context for interaction prediction.
- We propose an **optimal transport-based pseudo-labeling** strategy that treats label assignment as a point-set matching problem. This yields pseudo-labels aligned with the underlying data distribution and biological prior knowledge, improving the model's robustness and accuracy even with limited true labels.
- The **proposed framework achieves state-of-the-art performance** on multiple benchmark datasets, including virtual screening and MPI link prediction tasks. It outperforms prior approaches in terms of AUROC, early recognition metrics, and generalization to unseen interactions.

By bridging optimal transport with knowledge-driven representation learning, our approach provides a scalable and efficient solution for MPI prediction. Extensive experiments validate its effectiveness, particularly in leveraging multimodal biological signals to enhance interaction inference.

2 METHODOLOGY

2.1 OVERVIEW

We focus on the task of prediction of molecule-protein interaction, more specifically, mutual retrieval of molecule-protein, which involves two complementary objectives: (1) Given a molecule x, retrieve the protein capable of best catalyzing it. (2) Given a protein y, retrieve the molecules that can best bind to it.

To address this task, we collect a biological knowledge graph dataset that integrates molecular, protein, and knowledge graph (KG)-based methodologies. We then propose a novel framework that leverages optimal transportation (OT) to generate pseudo label of molecule-protein interactions. As shown in Figure 1 Our approach is designed as follows:

(1) Collection of knowledge graph dataset. We collected over 1.35 million interactions between molecules and proteins from UniProt and CHEBI datasets, and more interactions related to genes, genomes, protein families and enzyme comittee numbers. Please check Appendix B for details of our dataset.

(2) Training on a small labeled dataset: We use the sub dataset including only the molecule-protein pairs to train a scoring model that predicts the interaction strength for given pairs.

(3) Pseudo-label generation on a large unlabeled dataset: Using the trained model, we score molecule-protein pairs using all existing molecules and proteins in the dataset, these are mostly without pairwise labels and then we employ Optimal Transportation based method to assign high-quality pseudo-labels.

(4) Knowledge graph augmentation: We extract the predicted molecule-protein pseudo labels as supervision to the training on the full Knowledge Graph dataset for molecule-protein link predictions.

This framework enables the effective utilization of labeled and unlabeled data while leveraging relation information from KGs to enhance molecule-protein interaction predictions.

2.2 PREPARATIONS FOR PSEUDO LABEL GENERATION

Pretrained Backbone To represent molecules and proteins structural information effectively, we adopt pretrained encoders based on the Uni-MolZhou et al. (2023) framework as the backbone for extracting the features. Uni-Mol is a molecular and protein pretraining model specially designed to process molecule and protein 3D conformation data, and has achieved good performance on a variety of downstream tasks including molecule property prediction and bingding pose prediction

Both encoders produce embeddings $f(x)$ and $g(y)$ for molecules x and proteins y, respectively. These embeddings are normalized using the Euclidean norm to ensure consistency and compatibility for later tasks.

Molecular Similarity Calculation To measure the similarity between molecules, we utilize the embeddings extracted by the pretrained molecular encoder. Given two molecules x_i and x_j, their respective embeddings $f(x_i)$ and $f(x_j)$ are first computed using the molecular encoder.

The similarity between molecules x_i and x_j is then quantified using the cosine similarity, which is defined as:

$$\text{Sim}(x_i, x_j) = \frac{\langle f(x_i), f(x_j) \rangle}{\|f(x_i)\|_2 \|f(x_j)\|_2},$$

where $\langle f(x_i), f(x_j) \rangle$ is the dot product of the two embeddings, and $\|f(x_i)\|_2$ and $\|f(x_j)\|_2$ are their respective Euclidean norms.

This similarity measure serves as a foundation for incorporating molecular relationships into optimal transport-based pseudo label generation.

2.3 PSEUDO LABEL GENERATION WITH OPTIMAL TRANSPORTATION

In order to get the probability of interaction between certain molecule-protein pairs, we trained a score function to help the prediction. We first extract the features of the protein g (y) and the molecule f (x) using pre-trained encoders and then calculate the score S(x,y) for all the pairs of molecules and proteins, formulating a score matrix $S_{sup} \in R^{M*N}$ where $S_{sup}(i,j) = S(x_i, y_j)$. As inspired by Shi et al. (2023), the task of learning such relation can be represented as learning the transformation matrix $T \in R^{M*N}$ of the following optimal transport problem:

$$\min_{T \geq 0} \sum_{i=1}^{M} \sum_{j=1}^{N} T_{i,j} C_{i,j}, \quad \text{s.t. } T\mathbf{1}_N = r,\ T^\top \mathbf{1}_M = c, \tag{1}$$

Here, the cost matrix $C \in \mathbb{R}_+^{M \times N}$, and $C_{i,j} = 1 - S_{i,j}$ shows the cost for pairing. Source and sink distributions are defined as:

Source distrbution $r \in \mathbb{R}^M$ satisfies $r_i = \frac{1}{M}$; Sink distribution $c \in \mathbb{R}^N$ satisfies $c_j = \frac{1}{N}$;

$\mathbf{1}_M, \mathbf{1}_N$ are one dimension vectors with length of M and N.

Training Strategy of Score Function on Labeled Dataset To train the score function $S(x, y)$ effectively on a labeled dataset, we take an inverse optimal transportation perspective. The interaction score $S(x, y)$ between a molecule x and a protein y is modeled as a learned function of their embeddings:

$$S(x, y) = W(x \oplus y), \tag{2}$$

where W is a trainable model, and $\oplus$ represents concatenation.

During training, we construct ground truth cost matrices C_{gt} based on current batch of positive and negative molecule-protein pairs. For a given positive molecule-protein pair (x_i, y_i), the ground truth cost $C_{\text{gt}}(i, j)$ is defined such that:

$$C_{\text{gt}}(i,j) = \begin{cases} 0, & \text{if } j = i \text{ (positive pair)}, \\ 1, & \text{if } j \neq i \text{ (negative pair)}. \end{cases} \tag{3}$$

The theoretical optimal transport matrix T_{gt} is computed using C_{gt} as the cost matrix, following the Sinkhorn-Knopp algorithm to enforce marginal constraints:

$$T_{\text{gt}} = \arg\min_{T \geq 0} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{i,j} C_{\text{gt}}(i,j), \quad \text{s.t. } T\mathbf{1}_N = r,\ T^\top \mathbf{1}_N = c, \tag{4}$$

where r and c are uniform source and sink distributions, respectively.

For a batch of N molecule-protein pairs, we calculate the predicted transport matrix T_{pred} based on the cost matrix derived from the predicted scores $C_{\text{pred}}(i,j) = 1 - S(x_i, y_j)$. The loss function is defined as the KL divergence between T_{pred} and T_{gt}:

$$\mathcal{L}_{\text{score}} = \text{KL}(T_{\text{pred}} \| T_{\text{gt}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} T_{\text{pred}}(i,j) \log \frac{T_{\text{pred}}(i,j)}{T_{\text{gt}}(i,j)}. \tag{5}$$

This formulation ensures that the learned score function $S(x, y)$ aligns the predicted transport matrix with the theoretical optimal transport matrix derived from ground truth labels. As described in Shi et al. (2023), contrastive learning with the InfoNCE loss can be view as a special form of this method, we optimize the model to maximize the scores of true molecule-protein pairs while minimizing those of false pairs in the batch.

Pseudo label generation on large unlabeled dataset Let us consider the scenario where we have M molecules and N proteins. The pairing degree between molecules and proteins is represented as a matrix $S \in \mathbb{R}_+^{M \times N}$, where each element $S_{i,j}$ reflects the score between molecule i and protein j, calculated form the model we get in the former step. Our goal is to generate a pseudo label matrix $T \in \mathbb{R}_+^{M \times N}$ that can be used for further training. We also treat this problem as a Optimal Transport problem.

We define the optimal transport problem as follows:

(1) The *cost matrix* $C \in \mathbb{R}_+^{M \times N}$ is defined as $C_{i,j} = 1 - S_{i,j}$, where a smaller cost corresponds to a higher pairing degree.

(2) The *source distribution* $r \in \mathbb{R}^M$ represents the initial distribution over the molecules, with each entry given by $r_i = \frac{1}{M}$.

(3) The *target distribution* $c \in \mathbb{R}^N$ represents the distribution over the proteins, with each entry given by $c_j = \frac{1}{N}$.

The optimal transport problem is to find a transportation matrix $T \in \mathbb{R}_+^{M \times N}$ that minimizes the total cost while satisfying the marginal constraints:

$$\min_{T \geq 0} \sum_{i=1}^{M} \sum_{j=1}^{N} T_{i,j} C_{i,j}, \text{subject to } T\mathbf{1}_N = r,\ T^\top \mathbf{1}_M = c, \tag{6}$$

where $\mathbf{1}_M$ and $\mathbf{1}_N$ are uniform distributions over lengths M and N, respectively.

In our case, we have additional information about molecular similarity represented as a matrix $\text{Sim} \in \mathbb{R}_+^{M \times M}$, where $\text{Sim}_{i,k}$ quantifies the similarity between molecule i and molecule k. To leverage this information, we introduce an additional constraint: the similarity of pseudo labels between molecules i and k, denoted by $\text{Sim}_{i,k}^T = \sum_{j=1}^{N} T_{i,j} T_{k,j}$, should be as close as possible to $\text{Sim}_{i,k}$.

The modified objective function becomes:

$$\min_{T \geq 0} \sum_{i=1}^{M} \sum_{j=1}^{N} T_{i,j} C_{i,j} + \lambda \sum_{i=1}^{M} \sum_{k=1}^{M} \left(\text{Sim}_{i,k} - \text{Sim}_{i,k}^T\right)^2, \tag{7}$$

where $\lambda > 0$ is a weighting factor balancing the cost and similarity terms, we take $\lambda = 0.1$

As shown in the algorithm, we employ the Sinkhorn-Knopp algorithm to solve the optimal transport problem efficiently and extend it to handle the similarity constraints.

The Sinkhorn-Knopp algorithm solves the regularized optimal transport problem by introducing an entropic regularization term:

$$\min_{T \geq 0} \sum_{i=1}^{M} \sum_{j=1}^{N} T_{i,j} C_{i,j} + \epsilon \sum_{i=1}^{M} \sum_{j=1}^{N} T_{i,j} \log T_{i,j}. \tag{8}$$

The solution can be computed iteratively: (1) Initialize $u = \mathbf{1}_M$, $v = \mathbf{1}_N$, and $K = \exp(-C/\epsilon)$. (2) Iterate until convergence: $u \leftarrow \frac{r}{Kv}, v \leftarrow \frac{c}{K^\top u}$. (3) Compute T as: $T = \text{diag}(u) K \text{diag}(v)$.

Algorithm 1: Training Strategy on Large Unlabeled Dataset

Input: Pairing score matrix $S \in \mathbb{R}_+^{M \times N}$, molecular similarity matrix $\text{Sim} \in \mathbb{R}_+^{M \times M}$, source distribution $r \in \mathbb{R}^M$, target distribution $c \in \mathbb{R}^N$, entropic regularization parameter ϵ, similarity weight λ, learning rate η, maximum iterations max_iter.
Output: Optimal transport matrix $T \in \mathbb{R}_+^{M \times N}$.
Initialization:
Define cost matrix $C \in \mathbb{R}_+^{M \times N}$ as $C_{i,j} = 1 - S_{i,j}$.
Set $K = \exp(-C/\epsilon)$, $u = \mathbf{1}_M$, $v = \mathbf{1}_N$, and $T = 0$.
for $t = 1$ *to max_iter* **do**
  Step 1: Sinkhorn-Knopp Iteration.
  while *not converged* **do**
    Update $u \leftarrow \frac{r}{Kv}$.
    Update $v \leftarrow \frac{c}{K^\top u}$.
  Compute $T = \text{diag}(u) K \text{diag}(v)$.
  Step 2: Similarity Constraint Adjustment.
  Compute $\text{Sim}_{i,k}^T = \sum_{j=1}^{N} T_{i,j} T_{k,j}$ for all i, k.
  Compute gradient $\nabla_{T_{i,j}} = 2\lambda \sum_{k=1}^{M} \left(\text{Sim}_{i,k} - \text{Sim}_{i,k}^T\right) T_{k,j}$.
  Update $T \leftarrow T - \eta \nabla_T$.
  Step 3: Projection onto Feasible Set.
  Ensure $T \geq 0$, and normalize T such that $T\mathbf{1}_N = r$ and $T^\top \mathbf{1}_M = c$.
return T.

To incorporate similarity constraints, we modify T using gradient-based optimization. The gradient of the similarity term with respect to T is given by:

$$\nabla_{T_{i,j}} = 2\lambda \sum_{k=1}^{M} \left(\text{Sim}_{i,k} - \text{Sim}_{i,k}^T\right) T_{k,j}. \tag{9}$$

The algorithm alternates between updating T using the Sinkhorn steps and refining T with the similarity constraint gradient:

$$T \leftarrow T - \eta \nabla_T, \tag{10}$$

where $\eta > 0$ is the learning rate, and T is projected back to the feasible set if needed.

Pseudo Label Generation We generate the pseudo label matrix P, where each element P_{ij} represents the predicted likelihood of a catalytic interaction between molecule x_i and protein y_j. This matrix is computed using our multimodal score function $S(x, y)$. To ensure reliability, we extract the most confident interactions (x_i, y_j) with scores above a predefined threshold $\delta = 0.5$. These high-confidence pairs form a set of pseudo-labeled interactions, denoted as $\mathcal{P}$:

$$\mathcal{P} = \{(x_i, y_j) \mid S(x_i, y_j) > \delta\}. \tag{11}$$

These pseudo labels act as a source of augmented interaction data, compensating for the lack of large-scale ground truth molecular-protein interaction datasets.

2.4 Unified Framework for Link Prediction

To enhance the protein-molecule link prediction task, we propose a unified framework that integrates the pseudo label matrix T, generated by the optimal transport-based training strategy, with the structured knowledge from the knowledge graph (KG).

Table 1: Results on the DUD-E virtual screening benchmark (zero-shot setting). Higher is better for all metrics. Our OT + KG framework outperforms both traditional docking methods and modern learning-based approaches across all evaluation metrics.

Model	AUROC (%)	BEDROC (%)	EF@0.5%	EF@1%	EF@2%
Glide-SP (Halgren et al., 2004)	76.70	40.70	19.39	16.18	7.23
Vina (Trott and Olson, 2010)	71.60	–	9.13	7.32	4.44
NN-score (Durrant and McCammon, 2011)	68.30	12.20	4.16	4.02	3.12
RFscore (Ballester and Mitchell, 2010)	65.21	12.41	4.90	4.52	2.98
Pafnucy (Stepniewska-Dziubinska et al., 2017)	63.11	16.50	4.24	3.86	3.76
OnionNet (Zheng et al., 2019)	59.71	8.62	2.84	2.84	2.20
Planet (Zhang et al., 2023)	71.60	–	10.23	8.83	5.40
DrugCLIP (Jia et al., 2025)	80.93	50.52	38.07	31.89	10.66
KGOT	**83.45 ± 0.42**	**51.20 ± 0.35**	**39.10 ± 0.50**	**33.00 ± 0.47**	**11.20 ± 0.30**

Table 2: Results on the LIT-PCBA benchmark (zero-shot setting). Our method achieves the best performance across all metrics, illustrating its robustness on this more challenging dataset.

Model	AUROC (%)	BEDROC (%)	EF@0.5%	EF@1%	EF@5%
Surflex (Jain, 2003)	51.47	–	–	2.50	–
Glide-SP (Halgren et al., 2004)	53.15	4.00	3.17	3.41	2.01
Planet (Zhang et al., 2023)	57.31	–	4.64	3.87	2.43
Gnina (McNutt et al., 2021)	60.93	5.40	–	4.63	–
DeepDTA (Öztürk et al., 2018)	56.27	2.53	–	1.47	–
BigBind (Brocidiacono et al., 2023)	60.80	–	–	3.82	–
DrugCLIP (Jia et al., 2025)	57.17	6.23	8.56	5.51	2.27
KGOT	**62.45 ± 0.38**	**6.52 ± 0.22**	**9.12 ± 0.40**	**5.90 ± 0.28**	**2.50 ± 0.15**

The relations in KG dataset include all observed interactions in the KG as well as a new relation type, `pseudo_interaction`, which encodes the pseudo label scores T.

Multi-Objective Learning Framework Our model is optimized with a multi-objective loss that jointly leverages the knowledge graph structure and pseudo-label supervision. Specifically, we combine a graph embedding loss over KG triples with a pseudo-label alignment term that encourages predicted interaction scores to match the generated pseudo-label matrix. This joint formulation allows the model to balance structural knowledge with data-driven signals. Full mathematical definitions and implementation details are provided in Appendix F.

3 Experiments and Results

Our evaluation spans two settings that highlight different aspects of the proposed framework: (1) virtual screening benchmarks, which test the model's ability to retrieve active molecules for given protein targets in a zero-shot manner, and (2) knowledge graph link prediction, which tests the model's ability to identify held-out molecule–protein links using the integrated KG.

3.1 Evaluation on virtual screening tasks

Virtual screening benchmarks evaluate how well models can rank candidate molecules for a particular protein target, based on predicted binding likelihood. We consider two widely-used datasets: **DUD-E** and **LIT-PCBA**. We follow the evaluation protocol of recent works like DrugCLIP (Gao et al., 2023) to ensure fair comparison.

DUD-E benchmark. The DUD-E dataset (Mysinger et al., 2012) contains 102 protein targets, each with a set of known active molecules (binders) and a large set of decoys (inactive molecules). In total, DUD-E includes 22,886 active molecule–target pairs. We frame each target's screening as a ranking task: the model must assign higher scores to active molecules than to decoys. We evaluate performance using three metrics commonly used in virtual screening: (1) **AUROC** (area under the ROC curve), a threshold-independent measure of ranking quality; (2) **BEDROC** with $\alpha = 20$, which emphasizes early recognition of actives (important in screening scenarios); (3) **Enrichment Factor (EF)** at certain top fractions (e.g., EF@0.5%, 1%, 2%), which measures how many actives are found among the top-ranked subset compared to random expectation.

Leakage control. To preclude train–test leakage on the *molecule side*, we remove from the training pool any ligand whose Tanimoto similarity to *any* DUD-E test active is above a threshold of ≥ 0.60; we additionally report a stricter *Murcko-scaffold–out* variant in which all training ligands sharing the Bemis–Murcko scaffold with any test active are excluded (see Appx. §D). On the *protein side*, we exclude from training any protein whose sequence identity (computed by MMseqs2) to *any* DUD-E test target exceeds 60%; we also provide a *family–out* control by removing all training proteins mapped (via HMMER to Pfam-A) to the same families as DUD-E test targets. We use a single model (no ensembling or post-hoc re-ranking) for all runs.

Our model produces an interaction score for each molecule–target pair, which we use to rank molecules for each target. Table 1 summarizes the results on DUD-E in the zero-shot setting. We compare to classical docking methods (*Glide-SP* (Halgren et al., 2004), *AutoDock Vina* (Trott and Olson, 2010)) and learned baselines including similarity-based models (*NN-score* (Durrant and McCammon, 2011), *RFscore* (Ballester and Mitchell, 2010)), structure-based deep models (*Pafnucy* (Stepniewska-Dziubinska et al., 2017), *OnionNet* (Zheng et al., 2019)), and recent cross-modal representation models (*Planet* (Zhang et al., 2023), *DrugCLIP* (Jia et al., 2025)). Our method achieves the highest scores on all metrics. Notably, it outperforms the previous best model by a significant margin in early recognition metrics (e.g., BEDROC and EF), indicating that the OT-guided pseudo-labeling helps identify actives at the top of the ranking more effectively. In terms of AUROC, we obtain about 83.5%, which is an improvement of $\sim$2.5% over DrugCLIP (80.9%) and substantially higher than traditional docking (Glide SP: 76.7%). These results demonstrate the benefit of augmenting the limited training interactions with pseudo-labeled examples and KG-derived information.

LIT-PCBA benchmark. LIT-PCBA is another benchmark from the Therapeutic Data Commons, designed to be more challenging and address some biases in DUD-E. It comprises 15 protein targets with 7,844 experimentally confirmed actives and 407,381 inactives. The class imbalance is even more extreme here, making early retrieval metrics critical. We evaluate our model on LIT-PCBA under the same zero-shot setting.

Table 2 shows the results. We compare to several docking and deep learning baselines reported for this benchmark, including Surflex (Jain, 2003), Glide-SP, Gnina (McNutt et al., 2021), DeepDTA (Öztürk et al., 2018), BigBind (Brocidiacono et al., 2023), and DrugCLIP. Our framework again achieves the top performance on all metrics. In particular, we see improvements in AUROC (62.45% vs. 60.93% for the best baseline, Gnina) and in early enrichment (EF@0.5% of 9.12 vs. 8.56 for DrugCLIP). Although the absolute values are lower than DUD-E (due to LIT-PCBA's difficulty), the consistent gains indicate that our pseudo-label + KG approach generalizes well.

3.2 Molecule-protein link prediction task

We next evaluate our unified framework on the task of predicting molecule–protein links on knowledge graph. For this, we created a held-out set of known molecule–protein interactions from our collected KG data. Specifically, we withhold 60,000 molecule–protein pairs (edges) from the KG to serve as test examples, these pairs were not included in the training pseudo-label generation or KG training. Each test pair is a true interaction. The model must predict these links purely from the remaining training data in the KG and the pseudo-label enriched representations.

We cast this as a link prediction problem: given a molecule node, rank candidate protein nodes by the predicted likelihood of interaction (head entity prediction in KG terms). We evaluate using standard information retrieval metrics Hit@K: Hits@1, Hits@3, and Hits@5, which measure the proportion of test queries for which the correct partner appears in the top 1, top 3, or top 5 predictions, respectively. These metrics reflect the model's ability to place the true interaction at or near the top of the candidate list.

We compare our full model (which uses pseudo-labels and KG, as described) against ablated versions to quantify the effect of pseudo-labeling. In Table 3, we report results for several knowledge graph embedding models with and without our pseudo-label augmentation: PairRE, RotatE, MuRE, TorusE, and ComplEx-FF are five different KG embedding architectures. For each, we train one model using only the real KG edges (baseline) and another including the `pseudo_interaction` edges and the alignment loss. We observe that incorporating pseudo-labels leads to consistent improvements across all model types. The absolute gains vary by model, but the trend is clear:

the additional pseudo-labeled interactions help the KG embedding models better discriminate true links. Among the embedding methods, we found TorusE performed strongly, and with pseudo-labels it achieved the highest Hits@5 (74.9%). ComplEx-FF benefited remarkably from pseudo-labels (Hits@1 rising from 30.8% to 43.6%). These results highlight that our pseudo-labeling approach is model-agnostic and can enhance a range of link prediction algorithms by providing extra supervision.

Table 3: Knowledge graph link prediction performance (Hits@K) with and without KGOT generated pseudo-label augmentation.

Method	Hits@1	Hits@3	Hits@5
PairRE	10.0%	17.0%	21.4%
PairRE + KGOT	**10.9%**	**20.5%**	**26.4%**
RotatE	48.5%	61.6%	66.6%
RotatE + KGOT	**52.0%**	**63.9%**	**68.0%**
MuRE	**15.9%**	24.7%	31.0%
MuRE + KGOT	13.7%	**25.9%**	**31.2%**
TorusE	49.4%	64.2%	70.0%
TorusE + KGOT	**53.4%**	**65.2%**	**74.9%**
ComplEx-FF	30.8%	40.2%	44.4%
ComplEx-FF + KGOT	**43.6%**	**54.3%**	**58.6%**

The consistent improvements demonstrate that our pseudo-label generation successfully injects useful information into the KG. By effectively "filling in" likely interactions that were not explicitly in the KG, the model can learn from a more complete interaction network. This leads to better retrieval of held-out interactions, validating the core idea of our approach: leveraging unlabeled pairs through OT-based pseudo-labeling boosts prediction performance.

3.3 Ablation study.

We conduct ablation experiments to assess the contributions of our design choices, including (i) the use of OT loss versus standard InfoNCE contrastive loss, (ii) different pseudo-labeling strategies, and (iii) the integration of multiple biological knowledge sources. Across all settings, our proposed OT-based formulations and multimodal knowledge graph integration consistently yield higher link prediction accuracy. For example, OT loss improves over InfoNCE, OT+similarity pseudo-labeling achieves the best Hits@5, and adding GO, protein family, and pathway relations each provide incremental gains in Hits@1. Due to space limit, full details and complete results are reported in Appendix E.

4 Conclusion

In this work, we provide a unified biomedical knowledge graph model to tackle the challenge of molecular-protein interaction retrieval by integrating multi-modal biological data to improve molecule-protein interaction prediction and proposing a novel pseudo-label generation framework based on optimal transport (OT) to mitigate the scarcity of large scale labelled datasets. Our approach integrates multiple biological datasets into a unified knowledge graph, spanning drugs, proteins, genes, and biological processes, aligning predicted interaction distributions with the underlying graph structure,significantly improving retrieval performance across multiple benchmark datasets.

Extensive experiments validate the superiority of our approach, demonstrating consistent performance gains across various molecular-protein interaction prediction tasks.

Overall, our work presents a scalable and efficient framework for molecular-protein interaction retrieval, bridging the gap between structured biological knowledge and deep learning-based representation learning. We hope that our work can provide a paradigm for other data-lacking tasks in the field of computation biology, and ultimately point the way to using machine learning methods to build unified foundation models in the biological field.

5 RECOMMENDED ATTACHMENTS

5.1 ETHICS STATEMENT

This work uses publicly available biomedical resources to construct a multimodal knowledge graph and evaluate molecule–protein interaction prediction on established benchmarks (DUD-E, LIT-PCBA). No human subjects, PHI/PII, or clinical interventions are involved. We follow the licenses of all data providers.

5.2 REPRODUCIBILITY STATEMENT

We provide end-to-end details to enable reproduction: data sources, entity/edge schemas, and counts for the integrated KG (Appx. B); leakage-control splits and zero-shot protocols for DUD-E and LIT-PCBA (Sec. 3.1, Appx. D); the OT-based pseudo-label generation objective and algorithm (Sec. 2.3, Alg. 1), along with hyper-parameters and hardware specs (Appx. C). An anonymous repository with code and data is referenced in the appendix.

5.3 USAGE OF LLM

LLMs were not used to generate data, labels, or experimental results. The core methodology relies on pretrained molecular/protein encoders and an optimal-transport pseudo-labeling framework integrated with a biomedical KG. LLMs were used only as general-purpose assist tools for light copy-editing and minor code polishing; all technical ideas, algorithms, experiments, and analyses were conceived and implemented by the authors, who take full responsibility for accuracy and integrity.

REFERENCES

Mohammed AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC bioinformatics*, 20:1–10, 2019.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Amos Bairoch. The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305, 2000.

Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

Michael Brocidiacono, Paul Francoeur, Rishal Aggarwal, Konstantin I Popov, David Ryan Koes, and Alexander Tropsha. Bigbind: learning from nonstructural data for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 64(7):2488–2495, 2023.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*, 2023. doi: https://doi.org/10.1038/s41597-023-01960-3. URL https://www.nature.com/articles/s41597-023-01960-3.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 11 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36 (suppl_1):D344–D350, 10 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm791. URL https://doi.org/10.1093/nar/gkm791.

Alfredo Di Nola, Danilo Roccatano, and Herman JC Berendsen. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins: Structure, Function, and Bioinformatics*, 19(3): 174–182, 1994.

Jacob D. Durrant and J. Andrew McCammon. Nnscore 2.0: A neural-network receptor–ligand scoring function. *Journal of Chemical Information and Modeling*, 51(11):2897–2903, 2011. doi: 10.1021/ci2003889. URL https://doi.org/10.1021/ci2003889. PMID: 22017367.

Qizhe Fan, Xiaoqin Shen, Shihui Ying, and Shaoyi Du. Otclda: Optimal transport and contrastive learning for domain adaptive semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553, 2023.

Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. Pfam: the protein families database. *Nucleic acids research*, 42(D1):D222–D230, 2014.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44595–44614. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8bd31288ad8e9a31d519fdeede7ee47d-Paper-Conference.pdf.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Federico Gossi, Pushpak Pati, Panagiotis Chouvardas, Adriano Luca Martinelli, Marianna Kruithof-de Julio, and Maria Anna Rapsomaniki. Matching single cells across modalities with contrastive learning and optimal transport. *Briefings in bioinformatics*, 24(3):bbad130, 2023.

Siyuan Guo, Lexuan Wang, Chang Jin, Jinxian Wang, Han Peng, Huayang Shi, Wengen Li, Jihong Guan, and Shuigeng Zhou. M$\hat{3}$ -20m: A large-scale multi-modal molecule dataset for ai-driven drug design and discovery. *arXiv preprint arXiv:2412.06847*, 2024.

Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

Ajay N Jain. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry*, 46(4):499–511, 2003.

Yinjun Jia, Bowen Gao, Jiaxin Tan, Jiqing Zheng, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Liping Tan, Hongyi Cai, Yanwen Huang, Zhiheng Deng, Xiangwei Wu, Yue Jin, Yafei Yuan, Jiekang Tian, Wei He, Weiying Ma, Yaqin Zhang, Wei Zhang, Lei Liu, Chuangye Yan, and Yanyan Lan. Deep contrastive learning enables genome-wide virtual screening. *bioRxiv*, 2025. doi: 10.1101/2024.09.02.610777. URL https://www.biorxiv.org/content/early/2025/04/29/2024.09.02.610777.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. *arXiv preprint arXiv:2202.13196*, 2022.

Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, volume 380, pages 2739–2745, 2020.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012. doi: 10.1021/jm300687e. PMID: 22716043.

Maho Nakata, Tomomi Shimazaki, Masatomo Hashimoto, and Toshiyuki Maeda. Pubchemqc pm6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *Journal of Chemical Information and Modeling*, 60(12):5891–5899, 2020.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning*, pages 31408–31421. PMLR, 2023.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Marta M. Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Paweł Siedlecki. Pafnucy - a deep neural network for structure-based drug discovery. *ArXiv*, abs/1712.07042, 2017. URL https://api.semanticscholar.org/CorpusID:125295017.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46 (5):2699–2699, 2018.

Joana Vilela, Muhammad Asif, Ana Rita Marques, Joao Xavier Santos, Célia Rasga, Astrid Vicente, and Hugo Martiniano. Biomedical knowledge graph embeddings for personalized medicine: Predicting disease-gene associations. *Expert Systems*, 40(5):e13181, 2023.

Debby D Wang, Mengxu Zhu, and Hong Yan. Computationally predicting binding affinity in protein–ligand complexes: free energy-based simulations and machine learning-based scoring functions. *Briefings in Bioinformatics*, 22(3):bbaa107, 06 2020. ISSN 1477-4054. doi: 10.1093/bib/bbaa107. URL https://doi.org/10.1093/bib/bbaa107.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and Rishita Anubhai. Biobridge: Bridging biomedical foundation models via knowledge graph. *arXiv preprint arXiv:2310.03320*, 2023.

Ying Xia, Xiaoyong Pan, and Hong-Bin Shen. A comprehensive survey on protein-ligand binding site prediction. *Current Opinion in Structural Biology*, 86:102793, 2024.

Xiangying Zhang, Haotian Gao, Haojie Wang, Zhihang Chen, Zhe Zhang, Xinchong Chen, Yan Li, Yifei Qi, and Renxiao Wang. Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 64(7):2205–2220, 2023.

Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14):15956–15965, 2019.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *arXiv preprint*, 2023.

A Related Literature

A.1 Optimal Transport in Representation Learning

Optimal Transport (OT) is a mathematical framework originally developed to solve resource allocation problems by minimizing the cost of transporting mass from one distribution to another. Recently, it has been increasingly applied in machine learning and representation learning tasks due to its ability to compare distributions in a geometrically meaningful way. Unlike traditional distance metrics such as Euclidean or cosine distances, OT considers the geometry of the distributions, enabling it to capture fine-grained relationships between data points.

In the context of multimodal representation learning, OT has been used to align embeddings from different modalities by computing an optimal coupling between them. This approach ensures that structurally similar elements across modalities are closely matched, thus enhancing the quality of learned representations. Applications of OT have been particularly impactful in cross-modal tasks, such as image-text retrieval, molecular-protein interaction modeling, and domain adaptation. Notable works include the Sinkhorn-Knopp algorithmSinkhorn and Knopp (1967), which makes OT computationally efficient for large-scale datasets by introducing entropy regularization to the transport problem. Shi et al. (2023) aim to understand and generalize contrastive learning as a form of inverse optimal transport. The paper also shows that InfoNCE contrastive loss is a specific case of the proposed IOT loss. Other notable works in applying Optimal Transportation in feature learning include Fan et al. (2024), Lee et al. (2022) and Gossi et al. (2023)

A.2 Knowledge Graphs in Multi-Modality Molecular and Protein Tasks

Knowledge graphs (KGs) have emerged as powerful tools for integrating and modeling complex, heterogeneous data in multi-modality tasks, including molecular and protein-related studies. A KG represents entities (e.g., molecules, proteins, and biological pathways) as nodes and their relationships (e.g., binding, inhibition, or interaction) as edges, enabling the incorporation of prior biological knowledge into machine learning models.

In molecular and protein studies, KGs such as DrugBankKnox et al. (2024), ChEMBLGaulton et al. (2012), and STRINGMering et al. (2003) have been widely used to capture molecular-protein interactions and other biological relationships. By encoding domain-specific knowledge into graph structures, KGs enhance downstream tasks like molecular property predictionFang et al. (2023), drug-drug interaction predictionLin et al. (2020), and disease-gene association studiesVilela et al. (2023).

Recent advancements have also explored cross-modality tasks involving molecules and proteins by leveraging KGs as a common representation space. For example, BioBridgeWang et al. (2023) aligns molecule and protein embeddings through a knowledge graph and evaluates cross-domain retrieval tasks. Similarly, graph neural networks (GNNs) have been applied to propagate information across graph nodes, enabling the extraction of contextual embeddings that incorporate relational knowledge. Despite these successes, integrating KGs with multi-modality data remains challenging due to the heterogeneous nature of molecular and protein features, as well as the sparsity of some entity relationships.

B Details of dataset construction

To develop a comprehensive knowledge graph to study molecule and protein interactions, we considered 6 primary resources of biological and clinical information. The data resources provide widespread coverage of biomedical entities, including proteins, genes, drugs, molecules, biological processes, cellular components and protein families. These were high-quality datasets, either expertly curated annotations such as KEGG, widely-used standardized ontologies such as the Gene Ontology, or direct readouts of existing large scale unimodality dataset, such as CHEBI and UniProt.

(1)1.35 million potential catalytic activity relationships between over 30,000 molecules from CHEBIDegtyarenko et al. (2007) and 500,000 proteins from UniProtUniProt Consortium (2018).

(2)Genetic and genomic information from KEGGKanehisa and Goto (2000), capturing how molecular interactions are influenced by gene regulation and metabolic pathways.

(3)Functional annotations from Gene Ontology Ashburner et al. (2000), enriching molecular representations with biological process and cellular component insights.

(4)Protein family classifications from PFaMFinn et al. (2014), improving interaction predictions by incorporating shared functional domains.

(5)Enzyme Commission numbers from ENZYMEBairoch (2000), allowing for enzyme-substrate relationship modeling within our graph.

Head Entities		Tail Entities	
Type	Quantity	Type	Quantity
UNIPROT	5,956,325	GO	3,191,321
CHEBI	336,374	CHEBI	1,678,407
KEGG	92,184	PFAM	792,235
GO	89,235	KEGG_KO	407,307
EC	8,459	EC	304,428
KMODUL	1,275	KPATHWAY	89,989
		KCOMPOUND	16,695
		KMODULE	3,470

Table 4: Entities Distribution, where KMODUL represents KEGG_MODUL, KPATHWAY represents KEGG_PATHWAY, KCOMPOUND represents KEGG_COMPOUND, and KMODULE represnts KEGG_MODULE.

Our knowledge graph dataset is a multimodal knowledge graph with 8 types of nodes, 29 types of directed edges, 6,483,852 relationships between entities.

C Implementation Details

Backbone Architecture. We use Uni-Mol Zhou et al. (2023) for both molecular and protein encoders. Hidden dimension is set to 512 for both. The scoring MLP has two layers of size [512, 256, 1], with ReLU activations.

Labeled Dataset Training. Training on the labeled set uses a batch size of 128, learning rate 1e-4, and Adam optimizer with weight decay 0.01. The score function $S(x, y)$ is trained for 50 epochs with early stopping.

Pseudo-label Generation. For full-batch OT, we construct a score matrix for 10k molecules × 5k proteins. Sinkhorn $\epsilon = 0.01$, similarity weight $\lambda = 0.1$, learning rate $\eta = 1.0$, and 50 iterations. Top-k baseline selects the top 5 pseudo-labels per protein.

Knowledge Graph Training. KG embeddings are trained with embedding size 256 and margin 6.0. We use 1:1:1 sampling for real:negative:pseudo triples and a batch size of 1024. Pseudo-interactions are weighted via the loss term $\mathcal{L}_{\text{pseudo}}$.

Hardware. All experiments are run on 4 × NVIDIA A6000 48GB GPUs with 256GB RAM. Total training time for each variant is under 12 hours.

Code availablity Our code and data is available at: https://anonymous.4open.science/status/KGE-543D

D Leakage-Control Experiments

Protocol. To further mitigate train–test leakage beyond the default setup in the main text, we evaluate our model under a *strict* protocol that combines molecule- and protein-side filtering: (i)

on the **molecule side**, we adopt a *Murcko-scaffold–out* setting where all training ligands sharing the Bemis–Murcko scaffold with any test active are removed; (ii) on the **protein side**, we adopt a *family–out* setting by removing from the training pool all proteins that map (via HMMER to Pfam-A) to any family present among test targets. All other training, inference, and evaluation details (single model, zero-shot evaluation, fixed seeds) follow the main-text configuration.

Results on DUD-E (strict leakage control). Table 5 reports performance under the combined *Murcko-scaffold–out* (ligands) + *Pfam family–out* (proteins) protocol. As expected, absolute numbers are lower than in Table 1, yet early recognition remains strong under strict filtering.

Table 5: DUD-E under strict leakage control.

Model	AUROC (%)	BEDROC (%)	EF@0.5%	EF@1%	EF@2%
KGOT (strict)	81.78	51.04	38.91	32.47	10.35

Results on LIT-PCBA. Table 6 summarizes performance on LIT-PCBA under the same strict protocol. Given the greater class imbalance and challenge level of LIT-PCBA, we continue to emphasize early enrichment metrics.

Table 6: LIT-PCBA under strict leakage control

Model	AUROC (%)	BEDROC (%)	EF@0.5%	EF@1%	EF@5%
KGOT (strict)	61.22	5.96	8.92	5.34	2.27

E Ablation Experiments

E.1 OT Loss vs. Contrastive Loss (InfoNCE) in Supervised Training

We first compare the proposed optimal transport loss to a standard contrastive loss (InfoNCE) for supervising the scoring model. This experiment evaluates whether our OT-based loss offers an advantage over a more conventional pairwise contrastive approach.

Experimental Setup: We train the scoring model (MuRE-based encoder) on known entity pairs using either (i) our OT loss, or (ii) an InfoNCE loss. For InfoNCE, each positive pair is contrasted against multiple randomly sampled negative pairs (with temperature tuned to 0.1). All other training settings (learning rate, epochs, etc.) are kept identical. We evaluate the models on the link prediction task.

Following Table shows the performance comparison. The model trained with the OT loss achieves slightly higher accuracy than with InfoNCE. For instance, OT loss yields an MRR of 0.256 vs. 0.243 with InfoNCE, and Hits@10 of 45.1 vs. 42.7.

Training Loss	MRR	Hits@10
OT Loss (ours)	0.256	45.1%
InfoNCE Loss	0.243	42.7%

Table 7: Performance of the scoring model with OT-based loss vs. contrastive InfoNCE loss. OT loss yields a modest but consistent improvement in link prediction metrics.

The OT-trained model outperforms the InfoNCE variant, indicating that the OT formulation provides a beneficial supervisory signal. We attribute this gain to the OT loss's ability to consider the full distribution of true associations for each entity, rather than only one positive against negatives at a time. In knowledge graph link prediction, an entity can have multiple correct targets; the OT loss naturally accommodates multiple positive matches by optimizing a transport plan over them.

Table 8: Ablation on pseudo-label generation (illustrative Hits@5 on link prediction).

Strategy	Hits@5 (%)
No pseudo-label augmentation	68.5
Random pseudo-labels (matched count)	66.0
Top-k per protein ($k = 5$)	72.0
OT without similarity ($\lambda = 0$)	73.8
OT with high entropy ($\epsilon = 0.1$)	73.0
OT + similarity ($\lambda = 0.1,\ \epsilon = 0.01$)	**74.9**

E.2 Pseudo-labeling strategies

We compare no pseudo labels, random selection, top-k per protein, OT without similarity, high-entropy OT, and our full OT+similarity.

OT yields balanced pseudo labels with broader protein/molecule coverage than top-k, while the similarity prior filters implausible assignments. Excessive entropy (ϵ large) makes labels diffuse and less informative.

E.3 Components Ablation

To further analyze the contributions of different components, we perform an ablation on the knowledge graph variant of our model. Table 9 shows an ablation using the TorusE embedding method on the link prediction task, where we incrementally add sources of information. Starting with only molecule-protein edges (using the small labeled set, akin to a baseline without any additional knowledge), we then add gene ontology (GO) relations, protein family relations, and metabolic process (pathway) relations from the knowledge graph. Each addition yields an improvement in Hits@1. Specifically, incorporating GO relations (which provide gene-function context) boosts Hits@1 from 36.3% to 43.7%; adding protein family info further raises it to 47.4%; and including metabolic pathways brings it to 53.4%. This ablation highlights that each modality of biological knowledge contributes to better performance. It underscores the importance of multimodal data integration: the model with full knowledge (last row) performs significantly better than using only the labeled molecule-protein pairs.

Table 9: Ablation study on effect of integrating different knowledge graph relation types (using TorusE).

Model Configuration	Hits@1 (Link Prediction)
TorusE (molecules & proteins only)	36.3%
+ Gene Ontology relations	43.7%
+ Protein family relations (GO + PF)	47.4%
+ Metabolic process relations (GO + PF + KEGG + EC)	**53.4%**

F Multi-Objective Learning Framework

The framework is trained using a multi-objective loss function that integrates information from both the pseudo label matrix and the KG structure. The loss components are as follows:

Graph Embedding Loss A knowledge graph embedding model is trained to learn representations of entities and relations in the KG. The graph embedding loss is defined as:

$$\mathcal{L}_{\text{KG}} = \sum_{(h,r,t)\in\mathcal{D}_{\text{KG}}} \log\sigma(f(h,r,t)) \quad + \sum_{(h',r,t')\notin\mathcal{D}_{\text{KG}}} \log\sigma(-f(h',r,t')), \tag{12}$$

where $f(h,r,t)$ is the scoring function for a triple (h,r,t), σ is the sigmoid function, and $\mathcal{D}_{\text{KG}}$ and $\mathcal{D}'_{\text{KG}}$ are the sets of positive and negative samples, respectively.

Pseudo Label Alignment Loss The pseudo labels T represent predicted scores for protein-molecule interactions. We incorporate these into the model by defining a loss term that aligns the KG-predicted scores with T:

$$\mathcal{L}_{\text{pseudo}} = \sum_{i=1}^{M} \sum_{j=1}^{N} \left(f(e_i, r_{\text{pseudo}}, e_j) - T_{i,j} \right)^2 , \tag{13}$$

where e_i and e_j are the embeddings of molecule i and protein j, respectively, and r_{pseudo} is the learned embedding for the `pseudo_interaction` relation.

The total loss function for training the model is a weighted sum of the above components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KG}} + \alpha \mathcal{L}_{\text{pseudo}}, \tag{14}$$

where α is the hyperparameter that balance the contributions of the pseudo label and similarity terms, we take $\alpha = 0.1$ in the experiments.